Universe Scientific Publishing
Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

On Data Storage Technology based on Big Data

Shanshan RAO Zhengde BAO Chenxi CHEN
School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract

Big data storage process is an important link before data processing, with its efficient and accurate function of data storage. With the step of information science into, data from all walks of life all the time, exploding state, then more and more complex data types, the traditional relational database's ability to deal with the data already can not adapt to the development of the era of big data, a new storage technology is gradually replacing traditional database, this paper reviewed on the basis of the characteristics of big data the discussion of the mass data storage technology, the development trend of data storage are analyzed.

Key Words

Data Storage, Data Analysis, Cloud Computing

DOI:10.18686/jsjxt.v1i2.684

浅谈基于大数据的数据存储技术

饶姗姗 鲍正德 李晨曦
四川大学锦城学院计算机与软件学院，四川成都，611731

摘 要

大数据存储过程是数据处理之前的重要环节,以其高效、精准的存储数据的功能。随着信息科学的进不断步,各行各业无时无刻都在产生数据量,呈现爆炸式增长状态,随之数据类型越来越复杂,传统的关系型数据库的处理数据能力已经不能适应大数据时代的发展,新的存储技术逐渐取代传统数据库,本文基于大数据的的特性概述了对海量数据的存储技术的讨论,浅析了数据存储的发展趋势。

关键词

数据存储；数据分析；云计算

1.引言

随着信息数据的增加,并且类型多样,近几年,大数据开始走入人们的视线,伴随着云计算的出现,大数据更是迅速发展,目前大数据已经为人们熟知[1]。另一方面,大数据的发展加速了云计算的技术革新,为什么这样说呢,众所周知,大数据最明显的一个特点就是数据量庞大,要高效存储如此的数据量,若在传统的处理能力上是无法办到的,而基于云计算的数据库可以更好的解决这一问题;无论是国内还是国外,数据存储问题一直是备受关注,人们也在传统的数据库进行改进,至今用于数据存储处理的技术主要包括分布式、虚拟化等等技术,为了克服数据存储的问题.人们一直在不断探索.

2.大数据特性

大数据的到来时信息是信息时代的结果,什么是大数据,通俗的话大数据就是用一般的软件及技术无法处理的数据称为大数据,那么大数据具有那些特点呢？下面我们简单介绍大数据基本特性即 4V 特性：

2.1 数据量巨大 （Volume）

大数据最明显的一个特点即数据量庞大,从个人到企业再到国家,据统计,现如今产生的数据量已经从 TP 级别发展到 PB 的数据量,数据量庞大带来的存储问题是不可小觑的,那么我们就要根据不同的标准将庞大的数据量进行筛选,但是数据不是越多越好,我们要选择有

价值的数据,有一定结构的数据,对不符合人们预期希望的数据要进行数据清洗,这也是处理分析大数据最重要的环节。

2.2 数据类型复杂(Variety)

数据类型不单单是人们认为的数值类型,其中包括传统的结构化数据,除此之外还有非结构化的数据其中包括数字、图文等等一系列,准确的说一切可以描述物质特征的符号都可以称之为数据,数据分析就是要从这些数据中找到内部存在的联系,另一方面,如何才能更好的存储数据也是一个很难解决的问题。那么,数据类型如此复杂,数据存储成为制约大数据产业发展的原因之一。

2.3 价值密度低（Velocity）

数据呈现的价值可以是无价也可以是无意义的,比如著名的海燕系统,是现如今最严格的查交通违规的系统,每天监控的交通信息量庞大,但是快速的数据处理分析后的结果也许并不会给警方带来有效价值；另一个贴近我们生活的例子,航空公司的售票记录表面只是一些数字,但通过一定的数据分析可以找到内部的关联,为航空公司提供航班计划和售票政策。所以数据的价值需要通过一定分析方法并且以可视化的方法呈现给人们。

2.4 处理速度快

1 秒定律,由于信息发展迅速,据统计,人类一天将为产生大约 3 万亿字节的数据量,随着信息网络发展,这个数据量将会一直增加,不管是在生活中还是网络中,数据量都在不断的快速膨胀,如何应对这样的增大速度也成了人们面临的一大难题。但是不可否认的是,这个时代对信息发展的硬件和软件性能有更高的要求,从而才能适应信息时代的快速发展。

3.数据存储

随着数字图书数据、多媒体、电子商务企业、等数据的爆炸式增长,数据的容量单位从 TB 逐渐演变到 PB 量级,相应的数据存储容量也成了一大难题,数据存储除了对以往和现在数据量的保存,还要求对数据的更新以及防止数据快速膨胀而导致服务器负荷过重崩溃,所以,对海量数据的高效存储也是至关重要的[2]。

数据存储是指数据流在进行一定的活动或事务时需要查询的数据, 数据不是临时性的,不论是企业还是个人,往往都希望将数据保存下来,对于个人,历史数据不仅可以为当时还提供有效价值还能在今后帮助人们解决问题,数据存储有很多的途径,对于个人,数据量较少,一些基本的存储设备就可以满足用户需求,并且很容易实现和后期维护,比如 U 盘、电脑等方式,但是对于企业而言,数据量庞大并且要求保密性,不是 U 盘等存储设备可以满足的。

一般来说,数据是以某些特定的格式或规则存储在计算机内部,比如磁盘；此外还有一些外部存储设备,大数据的存储显然没有这么简单,大数据环境下的数据规模和复杂程度增加速度迅速,传统的数据库布能满足人类需要.需要研发新的技术,促进大数据研究领域对存储技术的研究。

4.大数据存储技术

大数据时代的关键不仅是帮助人们分析有价值的信息,更重要的时候如何将这些有价值的信息知识存储下来,为未来或当下提供有效的信息,大数据的出现同时伴随着信息产业的发展,促进存储技术的革新,面对数据量庞大、结构复杂的信息产业现状,应该采用什么样的方式来存储,也是信息行业为此一直努力探索。现如今应用的存储模型有：NoSQL、云计算存储、MPP、分布式,我们将重点讨论 NoSQL 数据库模型,同时分布式与传统模式在稳定性、效率、独立性方面也存在较大差异。

4.1 NoSQL

传统型关系型数据库在面对大数据时,数据存储和处理速度慢,扩展性和弹性较低,这也注定它们不能成为大数据存储的首要选择,因此 NoSQL 是为了满足信息产业需求逐渐发展而产生的数据管理技术,NoSQL 指一系列非关系型数据库,NoSQL 可以说是为大数据而生的,它打破了传统的数据库模型的局限性。

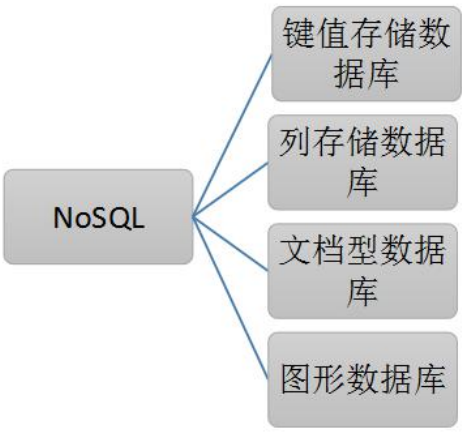
1) NoSQL 用于处理非结构化的数据类型,容纳复杂的数据模型(如图 1),相比传统数据库存储固定的数据结构扩展性更强和处理速度更快,例如：传统数据库每个元组的结构相同,即使某一实例无某种字段类型,也会被分配被定义的字段；而 NoSQL 模式是以键值对位标准,对数据进行存储,并没有首先对元组进行固定的格式化结构,而是根据不同元组得到不同需求而进行定义,

同时实现数据之间没有联系,从很大的程度上减少了空间资源的成本,以及时间的开销。

2) NoSQL 类型的数据库可以搭建在成本较低的硬件设备上,同时它也支持分布式存储,分布式存储是以网络环境,数据分布存储在不同的节点（服务器）,对用户而言,这样的工作方式是不可见的,这也让 NoSQL 具有高度扩展性、并且维护成本较低,下面我们介绍一个在 NoSQL 数据库技术中的两个成员-MongoDB 和 HBase

MongoDB 是一个面向文档的存储模式,是 NoSQL 的重要成员,它实质上介于关系型和非关系型的中间产物,可以支持存储多种复杂的数据类型,MongoDB 的文档一般都以 bso 的格式存储。

HBase 是列的存储架构模型,其原理是将不同列存储到列上,形成列簇,其中列簇即列的集合总称,可以更快的实时对数据列进行更新数据和读取。



4.2 分布式系统存储

4.2.1 分布式原理

数据类型多样,结构混合,处理起来非常的复杂.传统的数据库在数据增大到一定级别的时候,例如在处理十几个字段的数据表增加到几百个数据表,数据库的响应速度随之也会变得缓慢,其实这样的劣势与它的数据处理模式有关系,传统的处理模式为集中式存储方式,即非分布式处理。

数据存储在集中的服务器内,若其中某一台超负荷处理而崩溃,数据容易造成丢失。分布式系统存储是指将数据分成各个部分,让多台处理器并行处理各自数据,这一点类似于云存储的模式,云存储是将用户数据存储各个服务器上,有相应的监控机制和报警,某一节点崩溃不会导致其他节点同时崩溃[3]。分布式技术允许在一个时间点内多个合法用户访问存储数据和目录。另一方面分布式文件系统可以允许两个以上的节点同时执行相关数据库事务。

4.2.2 分布式与传统模式的比较

1)存取效率高

由于分布式处理模式采用的将数据分布在不同的处理器多台服务器,每台服务器事务的处理不用等待上一个事务完成,处理器间并行处理数据,例如 Hadoop-是一个分布式基础架构,就可以对大量数据进行分布式处理,若运行在它上面的处理器出来 1000MB 的数据,总共的存储时间由定位时间加上传输时间,若传输时间为 1s,调度为 2s,速度为 100MB/s,由 23 台机器同时传输,则需要 7.4 秒传输完成,而传统处理模式需要等待,等同串行传输数据,相比之下,分布式出来存储时间更少,效率自然也更高。

2)独立性、扩展性较强

由多台服务器同时工作,各部分相互独立当一台出现问题,并不影响其余服务器的进程任务,很大程度上提高了数据与系统的稳定性;分布式扩展性较强,分布式处理的各个数据放置在不同的地方,相互之间独立,当添加新的节点,不会影响数据的丢失,出现某一节点存储量过大,可以因而实现负载均衡,横向扩展性较强,而传统的模式则反之,它的处理模式决定了它弹性较差。

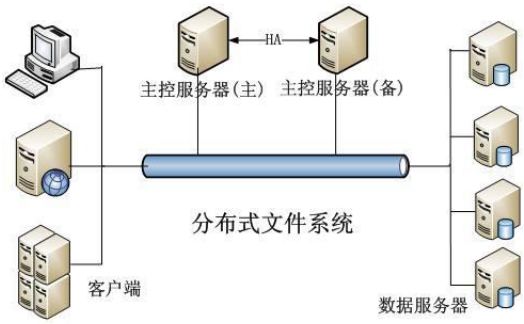


图 1

4.3 MPP 架构

MPP 架构是一种海量数据实时处理架构,主要运用在行业大数据,作为一种独享的架构,节点上运转自身的操作系统和数据,这种 MPP 架构目前被一些并行关系型数据库广泛运用 MPP 类的产品可以支撑的海量数据级

别为 TB,对于传统的数据数据库是不可想象的,所以对于数据仓库和结构化数据化的数据分析和存储处理,MPP 数据库是最佳选择。现如今采用这种技术的有查询系统 EMC、Google Dremel 等。

4.4 基于云计算数据库存储

云数据库是以云计算为基础,将数据库搭建在云计算环境下,,企业用户可以以租用的形式使用数据库资源,将数据存储在云数据库上,云数据库与传统的数据库相比较,云数据库采用虚拟化存储,虚拟化存储的关键是将物理设施等映射到逻辑的资源池,通过为客 3 户端和企业提供虚拟化存储空间,例如虚拟磁盘虚拟内存等,用户按照需要对资源自行组合,现如今云计算发展起来,云计算的三种形式 IAAS（基础设施即服务）作为云计算的基础,可以给用户提供存储空间,存储相应的数据,当然这只是满足普通用户的需求,用户数据存储处于在不同的服务器,若对较大的企业有海量的的数据可以自己搭建私有云平台,将企业数据存储在私有云,不仅以节省存储空间,优化存储虚拟化存储效率,同时减少存储成本。至今,云技术出现几年时间内快速发展,尤其是在云虚拟存储领域,如：网络云盘、虚拟机等应用。

5.总结

21 世纪是信息数据的时代,从生活到工作,人们无时无刻都在产生数据,而大数据带来不只是社会的改革,更是思维、科技的变革,本文从大数据特点到对大数据环境下数据存储做了简单的分析,我们不妨说,为了适应大数据时代的发展,面对不同的信息需求有相应的存储技术诞生,所以大数据带给我们的不仅是信息知识,更是引导存储技术变革方向,大数据时代已经拉开序幕,数据的存储技术需要不断的更新才能符合数据时代的要求,人类需要不断探索才能在数据存储技术上走的更远。

参考文献

[1] 朱孔村.大数据发展现状与未来发展趋势研究[J].大众科技,2019,21(01):115-118.
[2] 陈冠宇.大数据时代下计算机信息处理技术研究[J].网络安全技术与应用,2019(03):44+52.
[3] 陈帮鹏.“大数据”时代的计算机信息处理技术探讨[J].科技风,2019(08):95.
[4] 刘泉生.大数据时代计算机信息处理技术分析[J].科学技术创新,2019(05):82-83.
[5] 胡世昆.分布式数据库技术在大数据中的应用[J].电子技术与软件工程,2019(01):153.

作者简介

第一作者：饶姗姗（1997-）,女,汉,四川省成都市,本科,四川大学锦城学院,研究方向：电子商务。
第二作者（通讯作者）：鲍正德（1989-）,男,汉,黑龙江哈尔滨,研究生,四川大学锦城学院,研究方向：电子商务。
第三作者：李晨曦（1998-），男，汉，贵州省贵阳市，本科，四川大学锦城学院，研究方向：大数据技术开发